Making public use, synthetic files of longitudinal establishment data

Satkartar K. Kinney and Jerome P. Reiter [*]

Abstract

Longitudinal business data are widely desired by researchers, but difficult to make available to the public because of confidentiality constraints. In this paper, we discuss the generation of synthetic public use datasets for establishment data. The basic idea is to release simulated values of sensitive variables, generated from probability distributions fit using genuine data. This can protect confidentiality, since attributes are synthetic rather than real. And, when the models describe the data well, broad-scale inferences from the synthetic datasets will be inferentially valid. We illustrate approaches for generating synthetic establishment data by using LEHD infrastructure data.

1 Introduction

Many statistical agencies disseminate microdata, i.e., data on individual units, in public use files. These agencies strive to release files that are (i) safe from attacks by ill-intentioned data users seeking to learn respondents' identities or attributes, (ii) informative for a wide range of statistical analyses, and (iii) easy for users to analyze with standard statistical methods. Doing this well is a difficult task. The proliferation of publicly available databases, and improvements in record linkage technolgoies, have made disclosures a serious threat, to the point where most statistical agencies alter microdata before release. Some of the methods employed, such as data swapping (Dalenius and Reiss, 1982) or adding random noise (Fuller, 1993) reduce the utility of the released data and require complex procedures to analyze properly (Reiter, 2004).

One approach that does allow for valid inferences to be made using standard statistical methods is the use of multiple imputation to generate synthetic datasets. Rubin (1993) first proposed using multiple imputation to generate fully synthetic datasets for the purpose of statistical disclosure limitation. In this approach, the actual units and collected values from the confidential microdata are not released, only multiply-imputed datasets generated from models fit with the original survey data. Simple combining rules developed by Raghunathan *et al.* (2003) and Reiter (2005) allow users to make valid inferences using standard statistical methods and software. Fully synthetic data are further described in Rubin (1993), Raghunathan *et al.* (2003), and Reiter (2002).

While fully synthetic data represents a promising disclosure limitation approach, it is difficult to implement, and no agencies have yet used it to generate public use files. A variant, partially synthetic datasets, proposed by Little (1993), has been used successfully (Kennickell, 1997; Abowd

[*] Institute of Statistics and Decision Sciences, Box 90251, Duke University, Durham, NC 27708, USA

and Woodcock, 2001). Partially synthetic datasets retain the originally sampled units and some of their observed values while other values, such as key identifiers or sensitive values, are replaced with multiple imputations. Hence confidentiality can be protected while allowing valid inferences to be made. An advantage over fully synthetic datasets is that the inferences are less sensitive to the imputation models; however, some risks of disclosure remain. Combining rules for inferences with univariate and multivariate estimands from partially synthetic data have been developed by Reiter (2003, 2005). It is not expected that every possible analysis will be valid in the synthetic data. Users desiring complex or detailed inferences at the individual unit level may need to seek special access to the confidential microdata.

This paper describes current efforts to create partially synthetic datasets for longituding establishment data from the Longitudinal Employer-Household Dynamics program. In particular, entry and exit information, and annual payroll and employment data are desired. Real geographic and industry classification information are also included. The synthetic data generated should ideally preserve relationships among the different variables. Section 2 describes the methods used to generate synthetic data. Section 3 describes how these are being applied to LEHD. Section 4 discusses further work to be done. This work represents preliminary efforts to synthesize a longitudinal establishment dataset.

2 Synthesis Methods

We use the notation of Reiter (2003). Let $I_j = 1, j = 1, \ldots N$ indicate unit j was selected in the original survey. Let Y_{obs} be the $n \times p$ matrix of observed survey data for $I_j = 1$; let Y_{nobs} be the $(N - n) \times p$ matrix of unobserved survey data for the units with $I_j = 0$; and let $Y = (Y_{obs}, Y_{nobs})$. For simplicity we assume that all units respond fully to the survey, i.e., there are no missing values. We let X be the $N \times p$ matrix of design variables for all units in the population and assume that this is known. Synthetic datasets will be constructed based on the observed data $D = (X, Y_{obs}, I)$.

Partially synthetic datasets are constructed by replacing selected values in the observed data with r independent draws from posterior predictive distributions. Let $Z_j = 1$ indicate that unit j has been selected to have any observed values replaced with imputations. Let $Y_{rep}^{(i)}$ be the imputed values in the ith synthetic dataset, $i = 1, \ldots, r$, assumed to be drawn from $(Y_{rep}|D, Z)$, and Y_{nrep} be the unchanged values, which are the same in each dataset. Each of r synthetic datasets, $D^{(i)}$, is comprised of $(X, Y_{rep}^{(i)}, Y_{nrep}, I, Z)$. Imputations should only be made from the posterior predictive distribution of those units with $Z_j = 1$. We assume here that all units will have their values of confidential variables replaced, i.e., $Z_j = 1$ for all units.

When several variables are considered confidential, i.e., Y has dimension $N \times d$, specification of the joint posterior density $Y|X$ may be difficult. We write the joint distribution as a product of conditional densities. For $Y = (y_1, \ldots, y_d)$, sampling from $Y|X$ is thus achieved by sampling from $f(y_1|X)$, $f(y_2|y_1, X), \ldots, f(y_d|y_1, \ldots, y_{d-1}, X)$. This allows complex relationships to be modeled in a computationally feasible fashion. Some predictors can be omitted from the imputation model if an independence relationship is reasonable. For example, with longitudinal variables, values in year t may be assumed to be dependent on values in year $t - 1$, but not on values in previous years.

A similar approach was taken by Abowd and Woodcock (2004) for generation of longitudinal linked data, where observations are taken from multiple sampling frames. They approximated the joint density using a sequence of conditional densities defined by generalized linear models. A key difference is that each confidential variable y_k is drawn from the conditional distribution $f(y_k|X, Y_{-k})$ where Y_{-k} represents all confidential variables excluding y_k. While this may produce satisfactory results, it is not guaranteed to converge to a joint distribution.

Typical confidential variables for establishment data include highly skewed variables such as income, in which case normal linear models do a poor job of modelling the data. In addition, categorical variables common in establishment data such as geographic or industry indicators can have numerous categories which can make model fitting difficult. Below we describe a semiparametric approach for sampling from nonnormal data, an adaptation for binary or categorical responses, and a multinomial approach for categorical data.

2.1 Semiparametric Method

Generalized additive models (GAMs) (Hastie and Tibshirani, 1990) are flexible models that do not make strong assumptions about the relationships between variables, and thus are particularly useful when the relationships between variables are complex and poorly described by linear or generalized linear models. A non-parametric GAM is similar to a generalized linear model, with all linear coefficients replaced by smoothing functions. We use a semi-parametric GAM, which may contain both linear and smoothing components. A disadvantage of GAMs is that fitting can be computationally expensive for large datasets.

We would like to generate a synthetic variable $\tilde{y}_1$ for confidential variable y_1 by drawing from an appropriate conditional distribution $f(y_1|X)$. We approximate this by first fitting a GAM using the observed data to obtain predicted values $\hat{y}_{1i}(x_i), i = 1, \ldots, n$ and residual values r_{1i}. We then bin the residuals based on values of $\hat{y}_{1i}$ and sample with replacement within each bin to obtain new residuals $\tilde{r}_{1i}$. Thus $\tilde{y}_{1i} = \hat{y}_{1i}(x_i) + \tilde{r}_{1i}$.

To generate a synthetic $\tilde{y}_2$, we need to draw from $f(y_2|y_1, X)$. To approximate draws from this distribution, we obtain predicted $\hat{y}_{2i}(y_{1i}, x_i)$ using a GAM and save the coefficients so that we can determine $\hat{y}_{2i}(\tilde{y}_{1i}, x_i)$. We then bin the residuals based on values of $\hat{y}_{1i}$ and use the values of $\hat{y}_{2i}(\tilde{y}_{1i}, x_i)$ to determine which bin to sample from to obtain $\tilde{r}_{2i}$. Then $\tilde{y}_{2i} = \hat{y}_{2i}(\tilde{y}_{1i}, x_i) + \tilde{r}_{2i}$. Similarly, we can approximate the distributions $f(y_3|y_2, y_1, X), \ldots, f(y_d|y_1, \ldots, y_{d-1}, X)$.

The GAM is used in this case to preserve relationships amongst the variables while sampling from the error distribution provides perturbations for disclosure limitation. In a normal linear model, one would sample the posterior distributions of the coefficients and variance; however, in the case of the GAM, these are difficult to simulate. Simply replacing values with predictions from the GAM curve involves no random process and does not replicate the error distribution of points off of the curve; hence we have approximated this distribution. In Raghunathan (2003), values are replaced with predictions from a GAM; however, this is after resampling and refitting of the GAM to approximate redrawing the model parameters. A similar step could be added to our procedure, and should be in survey settings. Assuming for census data that the parameters are known, this step may be omitted. Further evaluation is needed to assess the impact on the propriety of the imputations when resampling the data. For simplicity, in the remainder of this paper, we do not draw new values of parameters before making imputations.

2.2 Binary and Categorical Responses

For binary and categorical responses without very many categories, we sample from binomial and multinomial distributions, using appropriate generalized linear models to obtain the sampling probabilities. Another approach for a binary response is to use the semiparametric approach to obtain synthetic values logit($\tilde{p}$), and then solving to obtain $\tilde{p}$. In our tests this was no better than this simpler approach.

To generate a synthetic variable $\tilde{y}_1$ for binary response y_1, we first fit a logistic model using the observed data to obtain predicted probabilities $\hat{p}_i(x_i), i = 1, \ldots, n$. The synthetic $\tilde{y}_{1i}$ is obtained by sampling from $Bin(1, \hat{p}_i(x_i))$. To approximate draws from $f(y_2|y_1, X)$ when y_2 is binary, we use the observed data to fit a logistic model to obtain $\hat{p}_i(x_i, y_{1i})$, and save the coefficients so that we may determine $\hat{p}_i(x_i, \tilde{y}_{1i})$. The synthetic $\tilde{y}_{2i}$ are then obtained by sampling from $Bin(1, \hat{p}_i(x_i, \tilde{y}_{1i}))$.

For categorical responses, the same approach is used, but a generalized logit model is used in place of a logistic model to obtain posterior probabilities $\hat{p}_{ij}(x_i), i = 1, \ldots, n; j = 1, \ldots, c$, where c is the number of categories in the response. A multinomial distribution is used in place the binomial.

2.3 Multinomial Method

When there are several categories in the response, or several categorical predictors, the generalized logit model can become impossible to fit. The multinomial-Dirichlet model provides a convenient framework for sampling from the posterior predictive distribution for a categorical y when X are categorical, and can also be used to impute missing data (Gelman *et al.*, 1995).

In the disclosure limitation setting, problems may arise when categories and categorical variables are too numerous. Let C be a unique category determined by categorical predictors in X and let y_C be the observed values of a categorical response variable corresponding to n_C units in C. If $n_C = 1$, or $y_{Ci}, i = 1, \ldots, n_C$ all have the same value, then the above procedures will impute synthetic values $\tilde{y}_C$ for y_C such that $\tilde{y}_C = y_C$. This creates a high risk of re-identification of y_C. Hence in our approach we add a positive probability that for a given category C, the $\tilde{y}_C$ generated may contain values not present in y_C.

3 Imputation of LEHD

The development of the methods described in Section 2 was motivated in part by the desire to generate public use files for longitudinal establishment data from the U. S. Census Longitudinal Employer-Household Dynamics program. Currently controlled access to this data is only granted to researchers by special agreement with the U. S. Census Bureau. The data that we are working with are described in Table 1. County and industry codes are not synthesized but all other variables must be synthesized for the data to be considered for public release. As this project is ongoing, the procedure described here may not correspond to any eventual public release files. We describe our current approach and provide a few summary tables to show some features of the observed data that are preserved in the synthetic versions.

Our strategy is to build up the joint distribution as described in Section 2. Here we describe our approach and present preliminary results for a subgroup defined by a 4-digit SIC, with approx-

Table 1: Variable Descriptions

Variable	Name	Type	Description
$x1$	County	categorical	Geographic Location
$x2$	SIC	categorical	Industry Code
$y1$	Firstyear	categorical	First Year Establishment is Observed
$y2$	Lastyear	categorical	Last Year Establishment is Observed
$y3$	Multiunit	categorical	Multiunit Status
$y4$	Employment	continuous	March 12 Employment (26 years)
$y5$	Payroll	continuous	Annual Payroll (26 years)

imately 24,000 units. We generate the synthetic datasets as follows:

1. Impute Firstyear using the multinomial approach of Section 2.3 to approximate draws from $f(y_1|x_1, x_2)$.

2. Impute Lastyear, treating Lastyear and Firstyear as continuous, using the semiparametric approach of Section 2.1 to approximate draws from $f(y_2|y_1, x_1, x_2)$.

3. Impute Multiunit status using generalized logit model as described in Section 2.2 to draw from $f(y_3|y_2, y_1, x_1, x_2)$.

4. Impute Employment and Payroll variables using the semiparametric approach to draw from $f(y_4^{(t)}|y_4^{(t-1)}, y_5^{(t-1)}, y_3, y_2, y_1, x_1, x_2)$ and $f(y_5^{(t)}|y_4^{(t)}, y_5^{(t-1)}, y_3, y_2, y_1, x_1, x_2)$, where t indicates a year between 1976 and 2001. It may also be desired to include $y_4^{(t-2)}, y_4^{(t-3)}$, etc.

The variable Firstyear contains 27 categories, namely the years 1975 through 2001. We predict this for one 4-digit SIC grouping using County and 6-digit SIC as predictors. There are over 3000 counties in the United States and numerous 6-digit SICs corresponding to a 4-digit SIC. This results a large number of unique county-SIC groups, that with 27 categories in the response, it is not always possible to use a generalized logit or similar model to predict the response. Furthermore, there are many county-SIC groups for which the observed unit(s) has only one observed value of Firstyear. Hence the multinomial approach of Section 2.3 was developed to handle this case. The marginal distribution is well-preserved using this approach as seen in Figure 1. Initial exploratory analysis suggests that conditonal relationships are also preserved.

To predict Lastyear from Firstyear, SIC, and County, the numerous unique Firstyear-SIC-County combinations with only one unit makes it difficult to model with even the multinomial approach. Lastyear can be approximated as a continuous variable and predicted using the semiparametric approach. As the last year of observed data is for 2001, there is a very large spike in the distribution of Lastyear in the year 2001, as seen in Figure 2. This distorted the relationship and thus a two-stage approach was used to impute the synthetic values. First, a logistic regression was used to predict whether or not Lastyear$=2001$, as described in Section 2.2. Conditional on Lastyear$\neq 2001$, the semiparametric approach was used to predict Lifetime, where Lifetime=Lastyear$-$Firstyear. Predicting Lifetime instead of Lastyear helped in the preservation

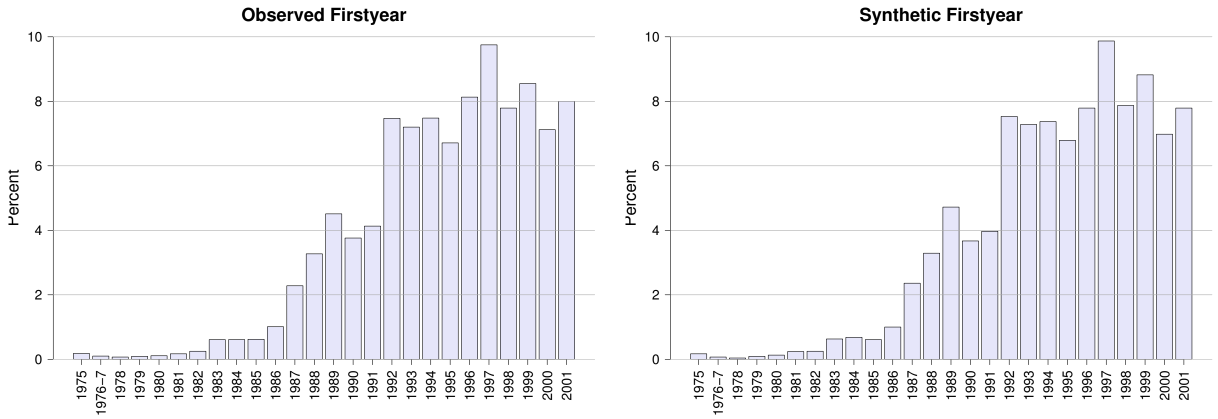

Figure 1: Observed and Synthetic Distributions of First Year

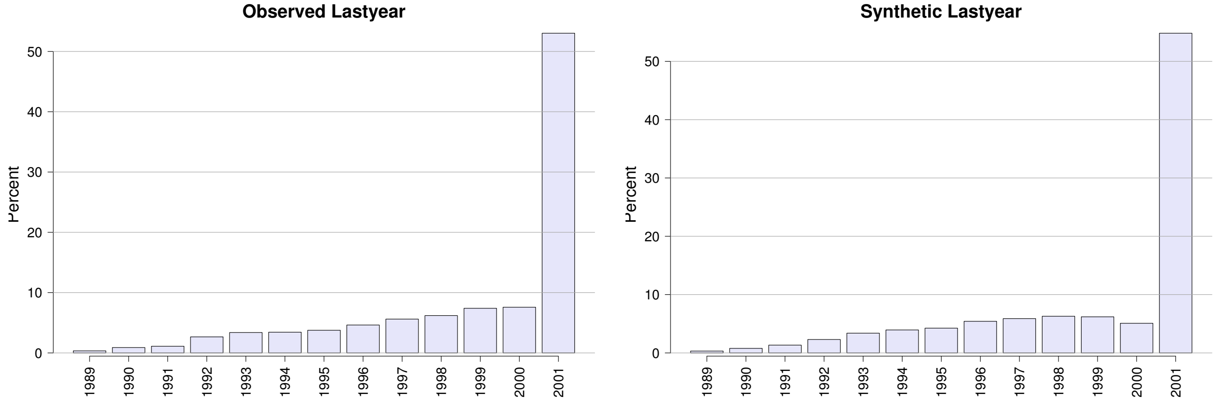

Figure 2: Observed and Synthetic Distributions of Last Year

of the distribution of Lifetime, along with Firstyear and Lastyear (Figures 1 to 3). One resampling step was applied to values with incongruous Lifetime and Lastyear values (e.g. Lastyear > 2001). Any remaining incongruous values were set to be the minimum or maximum value of observed Lastyear. As another check on the preservation of entry-exit information in the synthetic data, we also compare the observed and synthetic distributions of establishment age in 1990 (Figure 4).

The variable Multiunit indicates whether or not an establishment was ever part of a multi-unit firm. This variable was defined as a categorical variable with five values. A value of 1 indicates an establishment was never part of a multi-unit firm; values of 2-4 indicate a change in multi-unit status at some point in the lifetime of the establishment; and a value of 5 indicates the establishment was always part of a multi-unit firm. Synthesis of the categorical variable Multiunit using a generalized logit model was straightforward. Multiunit was derived from longitudinal binary variables indicating membership in a multi-unit firm. The time of change in status is also of interest and is planned for synthesis at a later time. Firm structure and linkages between establishments in the same firm are not likely to be synthesized. The marginal distributions of the observed and synthetic multiunit status are shown in Table 2.

Payroll and employment data are collected for each active establishment in every year between

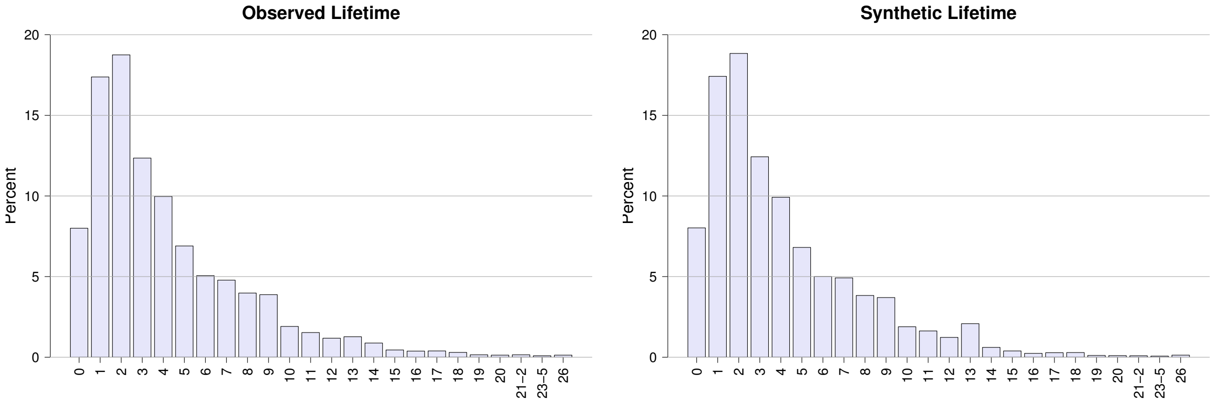

Figure 3: Observed and Synthetic Distributions of Lifetime

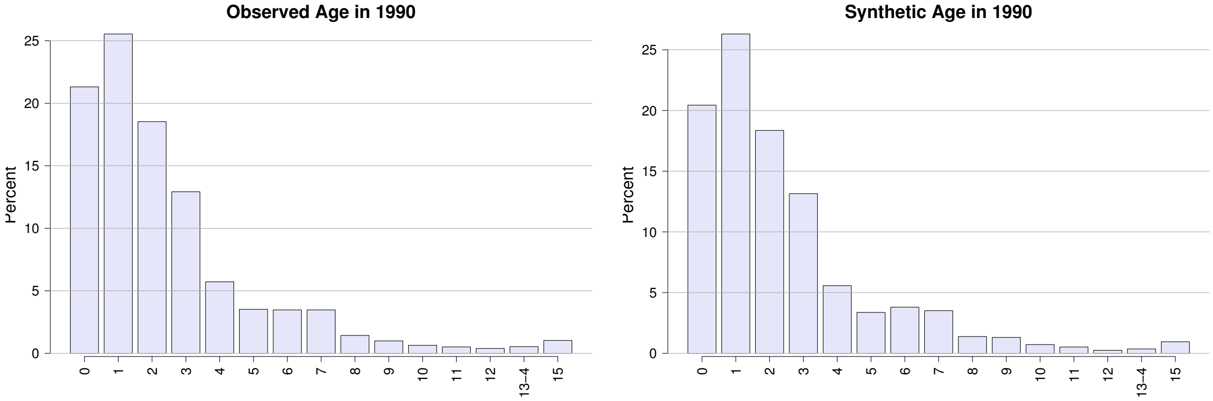

Figure 4: Observed and Synthetic Distributions of Age in 1990

Table 2: Observed and Synthetic Distributions of Multiunit Status

Value	Observed Percent	Synthetic Percent
1	84.08	83.86
2-4	0.96	0.90
5	14.97	15.25

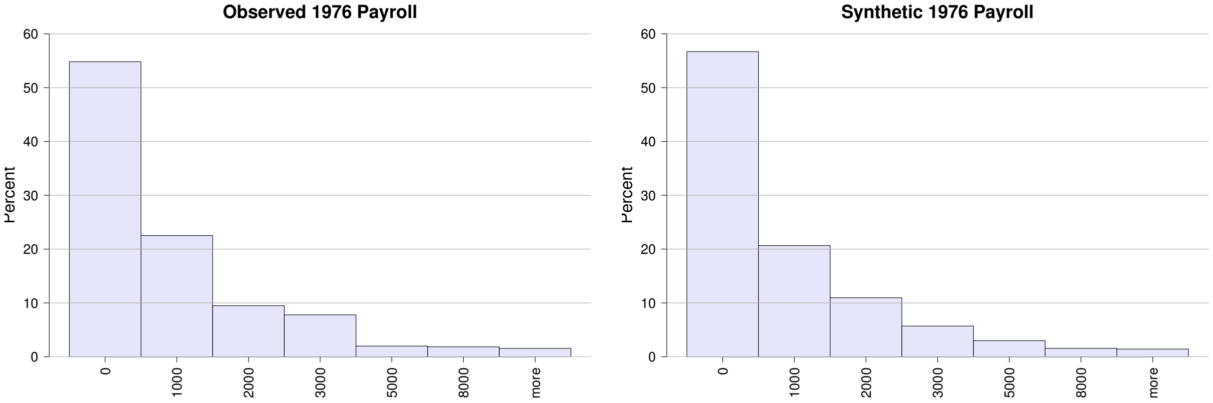

Figure 5: Observed and Synthetic Distributions of 1976 Payroll (in $1000)

1976 and 2001. If the synthetic values of Firstyear and Lastyear indicate an establishment was inactive in a given year, then no payroll or employment value is generated. For establishments in their first year, first-year employment and payroll are predicted from observed data corresponding to units in their first year. Payroll for continuers is predicted from the previous year's payroll, previous year's employment, as well as Multiunit, Firstyear, Lastyear, and State. Employment for continuers is predicted from previous year's employment, current year's payroll, and other predictors. Additional years of employment and payroll may be added as predictors. Initial results for this approach are promising; however, were not available in time to obtain disclosure approval for this writing. Summaries of 1976 payroll are shown in Figure 5 for a 4-digit SIC with around 1300 units. The mean and standard deviation for the observed data were 1167 and 2571 respectively, while for the synthetic data they were 1182 and 2766.

The goal of this project is the eventual release of public-use synthetic microdata. As an example of a disclosure check, we examine the observed values of Firstyear for units which had synthetic values of 1995 in three out of three synthetic datasets. As seen in Figure 6, there is a wide range of corresponding values in the observed data, suggesting a low risk of re-identification. A thorough evaluation will be completed prior to any data release.

4 Discussion

This paper has described ongoing work to develop synthetic data for longitudinal establishment data from the U. S. Census Bureau's Longitudinal Employer-Household Dynamics program for public release. The methodology is flexible and can be adapted for other datasets. Preliminary results illustrate the feasibility of using synthetic data for releasing microdata for public use while protecting confidentiality and allowing valid inferences to be made. When synthesis of the data has been completed further analysis of utility and risk will be completed before the data are released. Further evaluation of the randomization-validity of the proposed imputation methods is also needed.

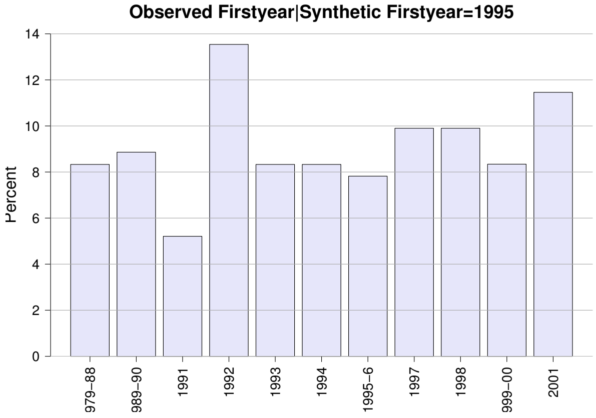

Figure 6: Observed Firstyear for Units with Synthetic Firstyear=1995

Acknowledgments

The research in this paper was conducted while the authors were Special Sworn Status researchers of the U.S. Census Bureau at the Triangle Census Research Data Center. Research results and conclusions expressed are those of the authors and do not necessarily reflect the views of the Census Bureau. This paper has been screened to insure that no confidential data are revealed. This work was supported by NSF grant ITR-0427889. The authors are also grateful for the assistance of many project participants and Census Bureau staff.

References

Abowd, J. M. and Woodcock, S. D. (2001). Disclosure limitation in longitudinal linked data. In P. Doyle, J. Lane, L. Zayatz, and J. Theeuwes, eds., *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, 215–277. Amsterdam: North-Holland.

Abowd, J. M. and Woodcock, S. D. (2004). Multiply-imputing confidential characteristics and file links in longitudinal linked data. In J. Domingo-Ferrer and V. Torra, eds., *Privacy in Statistical Databases*. New York: Springer-Verlag.

Dalenius, T. and Reiss, S. P. (1982). Data-swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference* **6**, 73–85.

Fuller, W. A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics* **9**, 383–406.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis*. London: Chapman & Hall.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. New York: Chapman & Hall.

Kennickell, A. B. (1997). Multiple imputation and disclosure protection: The case of the 1995 Survey of Consumer Finances. In W. Alvey and B. Jamerson, eds., *Record Linkage Techniques, 1997*, 248–267. National Academy Press, Washington, D.C.

Little, R. J. A. (1993). Statistical analysis of masked data. *Journal of Official Statistics* **9**, 407–426.

Raghunathan, T. E. (2003). Evaluation of Inferences from Multiple Synthetic Data Sets Created Using Semiparametric Approach. Tech. rep., National Academy of Sciences Panel on Access to Confidential Research Data.

Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics* **19**, 1–16.

Reiter, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics* **18**, 531–544.

Reiter, J. P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology* **29**, 181–189.

Reiter, J. P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology* **30**, 235–242.

Reiter, J. P. (2005). Significance tests for multi-component estimands from multiply-imputed, synthetic microdata. *Journal of Statistical Planning and Inference* **131**, 365–377.

Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics* **9**, 462–468.